〈BIGLOBEのインフラストラクチャ〉

“バーチャルデータセンタ”の実現に向けた技術開発

Technology of Virtual Data Center

遠藤由妃夫* Yukio Endo　小野雅弘* Masahiro Ono　石下隆一* Ryuichi Ishige
大石桐吾* Togo Oishi　岡崎洋平* Yohei Okazaki

要　旨

インターネットサービスの激しく変動するトラヒックの制御を実現し，サービスの可用性を向上するため，BIGLOBEでは「コンテンツ同期」と「トラヒック制御」の2つの技術を使用してデータセンタレベルまで仮想化したバーチャルデータセンタを構成しています。

BIGLOBE's Virtual Data Center is configured for the purpose of improving availability of Internet service and effective usage of resource under changeable traffic.

Synchronization of contents and control of traffic are necessary to the realization of BIGLOBE's Virtual Data Center.

1．まえがき

ブロードバンドの急速な普及により，インターネットは様々な場面において重要な役割を担うようになってきました。すでにインターネットは社会インフラであると言っても過言ではありません。

インターネットの利用者が増すにつれ，インターネット上で提供されるサービスが停止した場合の影響は大きく，予想される損失も膨大なものになっています。このため，以前にも増して安定したサービスの提供（可用性の向上）が求められるようになりました。

また，インターネットは今までのメディアにはない利点を備えており，情報配信メディアとして非常に注目を集めています。ブロードバンドの普及もあって配信可能な情報の自由度も増し，他のメディアと複合的にインターネットが利用されることも多くなりました。その結果，トラヒックの発生状況にも大きな変化が現れてきています。今までの一般的な個人利用者向けのコンテンツでは，通常時とピーク時のトラヒックの差はせいぜい数倍程度でした。しかし，テレビのような同報性の高いメディアと連動することにより，通常の十数倍，多いときでは数十倍のバーストトラヒックが発生することがあります。

このような可用性の向上，突発的なバーストトラヒックに対応するためのシステム，それを可能にするためのリソースの有効利用を従来のデータセンタで実現するには何点かの問題があります。

従来のデータセンタにおいても，各種装置の冗長化によって可用性を高めることは可能ですが，地震，洪水などの天災，あるいは火災などによるデータセンタ障害には耐えられません。また障害の復旧にも非常に長い時間を要してしまいます。

一方，突発的なバーストトラヒックに対応するためには，そのトラヒックを処理できるだけのリソースの確保が必須となります。しかし，バーストトラヒックに対応するために常に最大のリソースを準備するのは利用効率を考えると現実的ではありません。そのため現存する余剰リソースを利用することになりますが，データセンタに存在する余剰リソースには限界があり，要求を満たすだけのリソース確保が困難なことも多くあります。特にネットワークの帯域については各データセンタに余剰が少なく増速が必要な場合でも迅速に対応できないケースが出てきました。

そこで，これらの要求に応えるためにBIGLOBEでは，データセンタレベルの仮想化（バーチャリゼーション）に取り組んでいます。

2．バーチャルデータセンタ

バーチャルデータセンタ（**図1**）とは，複数のデータセンタを仮想的に1つのデータセンタとして構成するためのいくつかのキーとなる技術により実現されます。

そのキーとなる技術は大別すると「コンテンツ同期」と「トラヒック制御」の2つからなります。

「コンテンツ同期」は以下の3つの技術からなります。

＊ BIGLOBE構築運営本部
BIGLOBE Design and Operations Division

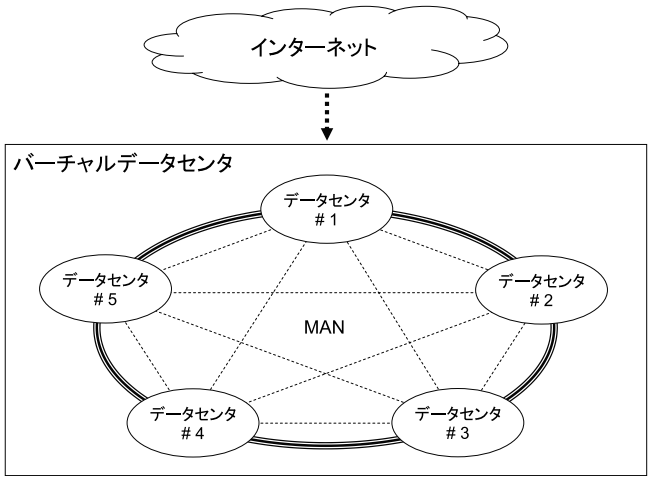


図1　バーチャルデータセンタ
Fig.1　Virtual Data Center.

・MAN（Metropolitan Area Network）
・ストレージ
・キャッシュサーバ

「トラヒック制御」は以下の3つの技術からなります。

・広域ロードバランサ
・コンテンツロードバランサ
・サーバ/スイッチ

バーチャルデータセンタでは，複数のデータセンタを仮想的に1つのデータセンタとして運用するために，まず複数のデータセンタへのリクエストの振り分けに広域ロードバランサを利用します。次に各データセンタに振り分けられたリクエストは，コンテンツロードバランサにより複数のサーバに振り分けられます。

また，各データセンタに振り分けられたリクエストすべてに対し，同一のコンテンツを返す必要がありますので，データセンタ間でコンテンツの同期を行う必要があります。このコンテンツの同期をMAN，ストレージ，キャッシュサーバの技術を組み合わせて実現しています。

こうしてデータセンタレベルまで仮想化することにより，1つのデータセンタで障害が発生した場合でも他のデータセンタを使用してサービスを継続できるようになり，BC/DR（Business Continuity/Disaster Recovery）を実現できます。さらに，各データセンタに存在する余剰リソースを組み合わせて利用することにより，リソースの有効利用，突発的なバーストトラヒックの処理が迅速かつ効率的に行えます。

3．コンテンツ同期

3.1　MAN

BIGLOBEではMANと呼ぶ数十Gのオーダで動作するスイッチングハブの集合体を，複数データセンタの中心に位置するようネットワークトポロジを構成しています。対外的なトラヒックで利用するMANとデータセンタ間で利用するMANをVLAN（Virtual Local Area Network）によって同一物理ネットワーク上に実現しています。これらのVLANはQoS（Quality of Service）の技術を使用し，サービストラヒックを優先させるよう制御した上で，余裕のあるときにはコンテンツデータのバックアップなどバッチ業務トラヒックを流し，リソースの有効活用を図っています。

このようにMANを利用することにより，多量のコンテンツでも，データセンタ間で容易にやりとりすることが可能となります。

3.2　ストレージ

コンテンツ同期実現のために，SAN（Storage Area Network），NAS（Network Attached Storage）の両環境において，データセンタをまたがってストレージ装置を共用できる構成にしています。

それぞれの環境で利用する仕組みは異なりますが，どのデータセンタにあるどの装置のコンテンツでも同じように利用できるよう，以下3つの仕組みを組み合わせてストレージの設計を行っています。

・専用MAN経由でサーバとストレージを接続し，データセンタ間でストレージを利用
・パケットごとの遅延を軽減するためにキャッシュ機能を利用しファイルアクセスのボトルネックを回避
・専用MANを利用したデータミラー環境を構築

上記の取り組みにより，各データセンタに点在する複数のストレージを，あたかもバーチャルデータセンタ上に存在する1つの巨大なストレージであるかのように見せています（**図2**）。

一方，ディザスタリカバリはデータセンタ障害と，ストレージ装置障害の2つを想定し，SAN，NASでそれぞれのサービスレベルに応じ，以下のようなデータ保全方式を採用しています。

①SAN型ストレージ
・データセンタ障害対策は，装置内のミラーデータを別データセンタに転送することにより実現

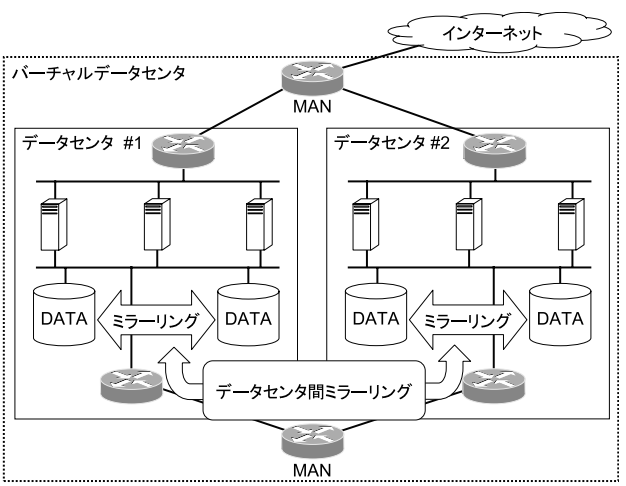


図2　ストレージの構成
Fig.2　Structure of storage.

・装置障害対策は別装置へのミラーリングにより実現

②NAS型ストレージ

・データセンタ障害対策は，snapshotを別センタに転送することにより実現

・装置障害対策はクラスタ内でのミラーリングにより実現

通常ディザスタリカバリを実現する場合，データのミラー先はディザスタリカバリセンタに存在するコールドスタンバイ装置となります。

BIGLOBEではストレージ内部の構成をサービス用とBC/DR 用に分離し，両方のストレージでサービスを行いながら，相互にディザスタリカバリセンタの役割を担う構成としています。これによってディスクだけでなく，CPU，キャッシュ，FC（Fibre Channel）インタフェースなどのヘッド部分も含めたストレージ全体をフル活用することが可能になります。

3.3 キャッシュサーバ

瞬間的なアクセス集中を複数のデータセンタで処理し，バーストトラヒックを緩和するためにキャッシュサーバを利用しています。

キャッシュサーバはユーザからのリクエストを受信すると，要求されたコンテンツが，自身が持つキャッシュ内に存在するか確認します。キャッシュを保持していればキャッシュしているコンテンツをリクエストに対し応答します。キャッシュを保持していない場合には，Webサーバに対してキャッシュサーバがリクエストを行い，取得したコンテンツをユーザに対し応答するとともに，キャッシュとして保持します。

最近のコンテンツはニュースや天気予報など短期間で更新するコンテンツが多いため，キャッシュしたコンテンツの鮮度確認が重要になります。鮮度確認の頻度を上げれば，確実に同期を取ることが可能ですが，その分Webサーバの負荷が増加してしまいます。

そこでBIGLOBEでは，以下の項目を判断材料として，キャッシュサーバの設計を行っています。

・サービスの特性

・コンテンツの更新頻度

・コンテンツの種類

・コンテンツの容量

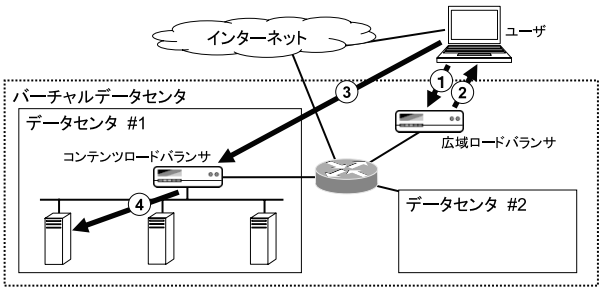
・発生するトラヒック量

これらの情報をあらかじめコンテンツごとに明確にし，キャッシュサーバの設計を行うことで，データセンタ間でのコンテンツ同期を正確に行いつつ，バーストトラヒック発生時にも遅延なくコンテンツの配信を行うことが可能となっています。

4．トラヒック制御

4.1 広域ロードバランサ

広域ロードバランサはDNSサーバとして動作します。



① ユーザからのDNSリクエストが広域ロードバランサに送信される

② 広域ロードバランサは各データセンタのリソース状況から最適なIPアドレスをユーザに通知する

③ ユーザは通知されたIPアドレスにアクセスする

④ コンテンツロードバランサにより最適なサーバが選択されユーザとのデータ交換が開始される

図3 トラヒック制御の流れ
Fig.3 Flow of traffic control.

ユーザからリクエストが発生する際，リクエスト先のホスト名はDNSサーバによってIPアドレスに変換されます。広域ロードバランサは，現在のサービスサーバの稼働状況，負荷状況を監視し，最適なデータセンタのIPアドレスを通知します（**図3**）。

広域ロードバランサが監視できる項目は，コネクション数，トラヒック量，応答時間など多岐にわたります。これらの情報を元に通知するIPアドレスを動的に変えることで，データセンタをまたがった負荷分散を行っています。

また，データセンタのリソースを考慮して，トラヒック量やコネクション数による制限を設けています。

広域ロードバランサの利用により，各データセンタへのトラヒックを効果的に制御し，リソースの利用効率を飛躍的に高めるとともに，データセンタ障害時のディザスタリカバリを自動で行えるように設計しています。

このような効果を十分生かせるようBIGLOBEでは以下のようなサービスで広域ロードバランサを積極的に利用しています。

・提供期間が限定され，1データセンタでは処理できない高帯域を必要とするストリーム系サービス

・テレビ番組と連携し，一時的にバーストトラヒックが予想されるイベント系サービス

・高可用性の求められるサービス

4.2 コンテンツロードバランサ

コンテンツロードバランサは，一般にはロードバランサあるいはL7SW（Layer 7 Switch）などと呼ばれますが，ここでは広域ロードバランサと明確に区別するため，コンテンツロードバランサと呼称します。

コンテンツロードバランサは，自身が受信したリクエストをデータセンタ内にあるサーバ群に振り分けます。バーチャルデータセンタでは，コンテンツロードバランサへのリクエストは，広域ロードバランサから誘導されたものであり，広域ロードバランサはデータセンタ間の負荷分散を，

コンテンツロードバランサはデータセンタ内の負荷分散を行うという違いがあります。

コンテンツロードバランサは，広域ロードバランサの監視項目よりも詳細にサービスサーバの稼働状況や負荷状況を監視することができ，リクエストに含まれるアプリケーション層のパラメータを参照することも可能です。

また，コンテンツロードバランサは，メンテナンス中のサーバにリクエストが振り分けられないよう，意図的にトラヒックを制御することも可能です。

BIGLOBEではコンテンツロードバランサを全サービスで利用し，高度な負荷分散機能によるリソースの利用効率の向上だけでなく，サービスを停止することなく各サーバのメンテナンスを可能とし，可用性の飛躍的な向上を実現しています。

4.3 サーバ/スイッチ

BIGLOBEのネットワーク機器は，ルータ，レイヤ3スイッチ，コンテンツロードバランサ，スイッチ，サーバという形で多段構成をとっています。

バーチャルデータセンタの構築に当たり，スイッチ，サーバを含め，ほぼすべての機器でTAG-VLANを利用してネットワークを構成しています。

TAG-VLANを用いることにより，サーバ，スイッチの1つの物理配線上に複数の仮想的な配線が設定可能となるため，より柔軟なネットワーク設計が可能となり，コンテンツロードバランサとの連携によりリソースの利用効率の向上に寄与しています。さらに，配線コストの削減，物理的な配線故障の監視簡略化にもつながっています。

またBIGLOBEで使用するブレードサーバに関しては，インストールから設定までをすべてリモートから行えるシステムを開発しており，現地作業などのデータセンタを意識した作業は必要ありません。

5．システム構築

ここまで説明してきたバーチャルデータセンタの1つの構築例として実際のサービス構成について紹介します。

このサービスでは，前述したバーチャルデータセンタのキーとなる技術のうち以下のものを用いて構築されています。

・MAN
・キャッシュサーバ
・広域ロードバランサ
・コンテンツロードバランサ
・サーバ/スイッチ

このサービスでは，提供されるコンテンツは通常の静的コンテンツと，3種類のビットレートを持つ動画からなります。そして雑誌，テレビなどの他のメディアとも連動していました。このためテレビCM放送時の瞬間的なバーストトラヒックは非常に大きくなると予想されました。しかしながらシステムの公開期間は限定的なもので，新規にこのバーストトラヒックに対応するだけのシステムを構築する投資は考えられませんでした。

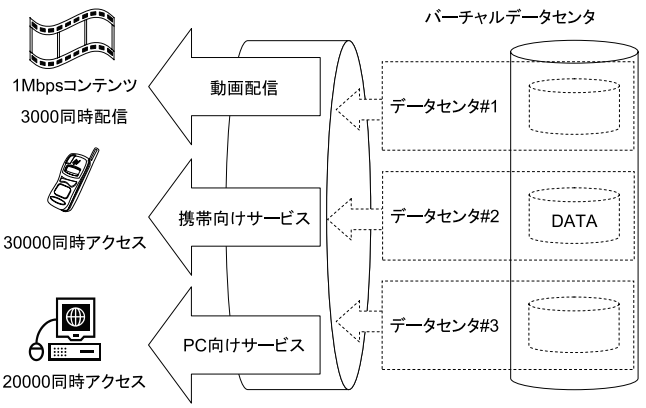


図4　バーチャルデータセンタの効果
Fig.4　Effect of Virtual Date Center.

まず，コンテンツを提供するためのWebサーバを数台構築しましたが，このWebサーバだけでは予想されるバーストトラヒックを処理するには不十分でした。そこで各データセンタにあるキャッシュサーバのうち，比較的負荷の低いものを集め，このWebサーバ上のコンテンツをキャッシュさせました。そして同一のデータセンタ内のキャッシュサーバをコンテンツロードバランサにより負荷分散し，各データセンタのコンテンツロードバランサに対しては広域ロードバランサにて負荷分散を行うようにシステムを構築しました。

上記の構成により，3Gbpsを超えるストリーミング配信能力を迅速に確保でき，テレビCM放送によるバーストトラヒック発生時も遅延などはなく，最低限のコストでサービスを提供することができました。

また，テレビと連動した他のサービスではピーク時の20,000を超える同時コネクションを問題なく処理し，その有効性が確認できました（**図4**）。

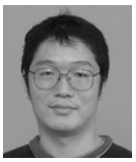
6．むすび

BIGLOBEでは，サーバやストレージのリソースを必要なときに必要なだけ利用できるグリッド技術を用いたシステム化を進めています。さらに，サービスのシステム設定と空リソースの判断，動的再配置を自動化するために自律コンピューティング技術開発を進めています。

BIGLOBEのバーチャルデータセンタは，グリッドおよび自律コンピューティングとしてさらなる進化をし，リソース利用効率の極大化，可用性，拡張性のさらなる向上，より質の高いサービスをより安くお客様に提供ができることと信じています。

筆者紹介

Yukio Endo
遠藤由妃夫 1986年，NEC入社。現在，BIGLOBE事業本部BIGLOBE構築運営本部グループマネージャー。



Masahiro Ono
小野　雅弘 1991年，NEC入社。現在，BIGLOBE事業本部BIGLOBE構築運営本部マネージャー。

Ryuichi Ishige
石下　隆一 1992年，NEC入社。現在，BIGLOBE事業本部 BIGLOBE構築運営本部主任。

Togo Oishi
大石　桐吾 1993年，NEC入社。現在，BIGLOBE事業本部 BIGLOBE構築運営本部主任。

Yohei Okazaki
岡崎　洋平 2001年，NEC入社。現在，BIGLOBE事業本部 BIGLOBE構築運営本部勤務。